Detecting Fake Hotel Reviews

using:
- Natural Language Processing
- Topic Modeling
- Machine Classification

SpringBoard – Capstone 2

11/2/2017

Chris Gian

github.com/chrisgian

Roadmap

- Motivation
- Data
- Data Preparation
- Topic Modeling
- Classification and Tuning
- Results
- Future steps

Motivation (1): Recent news

KOTAKU

Assassin's Creed Origins Metacritic Flooded With Fake Positive User Reviews

Jason Schreier
Yesterday 10:15am · Filed to: METACRITIC

159.9K 221 15

INVESTIGATIONS

IN-DEPTH INVESTIGATIVE REPORTING FROM NBC STATIONS ACROSS THE COUNTRY

Yelp, Facebook, Google Remove Fake Reviews

Yelp, Google and Facebook all suggest that consumers notify them of suspicious reviews.

HUFFPOST

ENTERTAINMENT 08/17/2016 06:10 pm ET

The Worst Of The Internet Is Sabotaging Amy Schumer's Book With Fake 1-Star Reviews

Can we not?

Tim Cook just outlined Apple's position on 'fake news'

Kif Leswing
17h 3,790

FACEBOOK | LINKEDIN | TWITTER | EMAIL | PRINT

BUSINESS INSIDER

Apple CEO Tim Cook. REUTERS/Toru Hanai

Motivation (2) Why + Approach

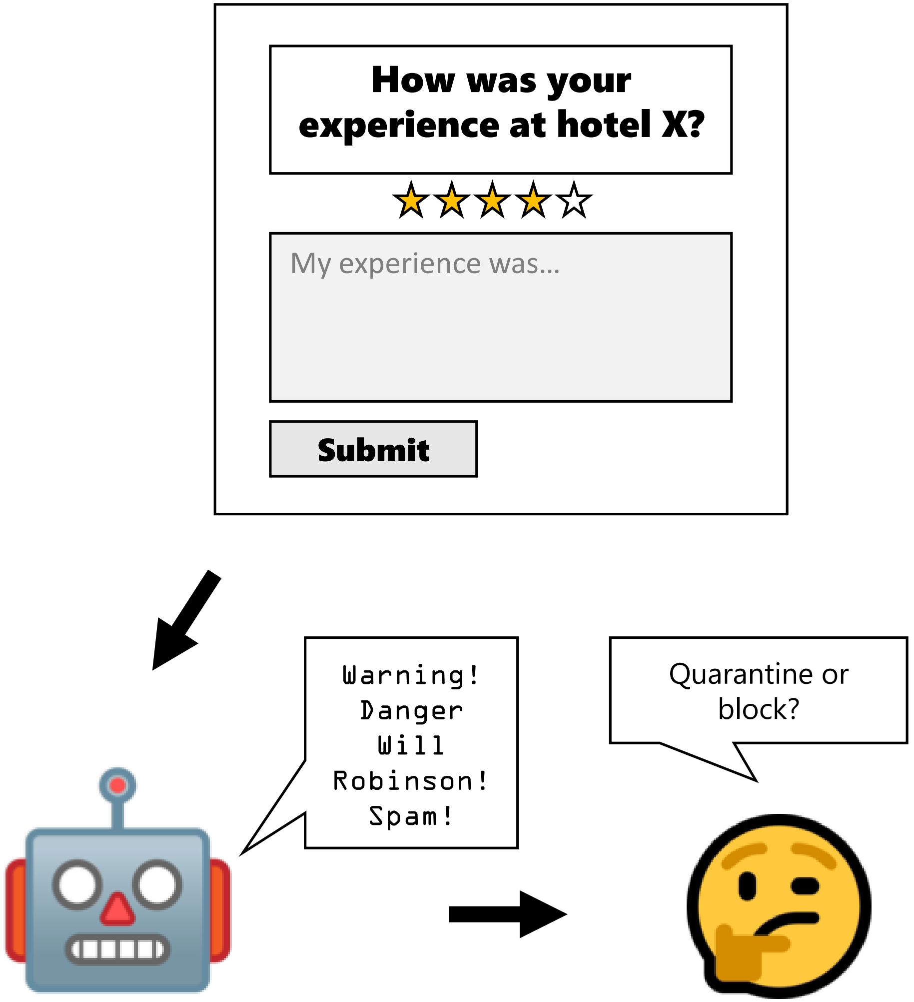
- Why is this important to me?
- Natural Language Processing (computational linguistics)
  - To extract, interpret, and translate information from one language into a form that can be stored, indexed, searched, and acted upon.

Motivation (3) Hypothetical Client

Client:
A Hotel Booking site.

Problem:
Users need reliable and honest hotel review information.

Solution:
Use **NLP** + **Topic Modeling** + **Classification** model to train a Spam Classifier



Motivation (4)
Can you tell the difference?

	Paid Reviewer	Real Reviewer
+	the experince at the hard rock hotel in chicago was fantastic,i will rate them a 6 out of 5. they have wonderful service and great staff and the view is just wonderful.	I recently stayed at the Hard Rock Hotel in Chicago, Il. From the start, the experience was bad. The room was filthy, there were no towels, and the front desk did nothing to rectify the situation. I will never stay there again. I could not have been more dissatisfied.
-	The Swissotel Chicago is a very mediocre hotel, the service is always poor, and the room service food always comes cold, unless it's supposed to be cold than it comes warm. I would rather stay at a super 8 than this place again.	Uhhhhh, how do I know which direction to go in??? Stains on the carpet in the room. A big gauge in the wall, where maybe a thermostat once was??? The shower is decent. All in all, for the price they charge, NO THANKS!!! I'm just glad my company is footing the bill.

Data (2) Descriptives

A very balanced data set

	source				text			
deceptive	deceptive		truthful		deceptive		truthful	
polarity	negative	positive	negative	positive	negative	positive	negative	positive
hotel								
affinia	20	20	20	20	20	20	20	20
allegro	20	20	20	20	20	20	20	20
amalfi	20	20	20	20	20	20	20	20
ambassador	20	20	20	20	20	20	20	20
conrad	20	20	20	20	20	20	20	20
fairmont	20	20	20	20	20	20	20	20
hardrock	20	20	20	20	20	20	20	20
hilton	20	20	20	20	20	20	20	20
homewood	20	20	20	20	20	20	20	20
hyatt	20	20	20	20	20	20	20	20
intercontinental	20	20	20	20	20	20	20	20
james	20	20	20	20	20	20	20	20
knickerbocker	20	20	20	20	20	20	20	20
monaco	20	20	20	20	20	20	20	20
omni	20	20	20	20	20	20	20	20
palmer	20	20	20	20	20	20	20	20
sheraton	20	20	20	20	20	20	20	20
sofitel	20	20	20	20	20	20	20	20
swissotel	20	20	20	20	20	20	20	20
talbott	20	20	20	20	20	20	20	20

Data Preparation: Tokenizing

Sentence → Tokens → Parts of Speech

1 text	2 tokens	3 tokens_stopwords	4 lemma	5 pos
We stayed for a one night getaway with family ...	[We, stayed, for, a, one, night, getaway, with...	[We, stayed, one, night, getaway, family, thur...	[-PRON-, stay, for, a, one, night, getaway, wi...	[PRON, VERB, ADP, DET, NUM, NOUN, NOUN, ADP, N...

Parts of Speech → <u>4 Parts of Speech Variables</u>

pos
[PRON, VERB, ADP, DET, NUM, NOUN, NOUN, ADP, N...

pron_ct	noun_ct	punct_ct	verb_ct
5	30	12	14

Data Preparation: Categorical Data

Categorical Data → Dummy Variables <u>(22 Variables)</u>

hotel	polarity	source
conrad	positive	TripAdvisor
hyatt	positive	TripAdvisor
hyatt	positive	TripAdvisor
omni	positive	TripAdvisor
hyatt	positive	TripAdvisor

hotel_palmer	hotel_sheraton	hotel_sofitel	hotel_swissotel	hotel_talbott	polarity_negative	polarity_positive
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1

Topic Modeling

- Latent Semantic Analysis:
  - Transform words used by reviewers into "groupings" – these are "topics"
  - Examples:
    - "Great family vacation", "Terrible room service", "No pets allowed"

	I	We	Family	Hotel	Chicago	…	Word M
Review 1	1	0	0	0	0	…	0
Review 2	1	1	1	0	0	…	1
Review 3	1	0	0	1	0	…	1
Review 4	0	1	0	0	1	…	0
Review 5	0	1	0	1	1	…	0
Review 6	1	0	1	0	0	…	0
Review 7	1	1	0	0	0	…	0
Review 8	1	1	0	1	1	…	0
…	…	…	…	…	…	…	0
Review N	1	1	0	0	1	0	1

Reduce into x "Topics"

	Topic 1	Topic 2	Topic 3	…	Topic M
Review 1	0.816027	0	-0.71844	…	0.678083
Review 2	0	0	0	…	0
Review 3	0.145131	0	0.533187	…	0
Review 4	-0.45718	0.238589	0.294818	…	0
Review 5	-0.08775	-0.29514	-0.18806	…	0
Review 6	0	-0.05889	-0.34565	…	-0.37266
Review 7	0	0	0	…	-0.1482
Review 8	0	0.763458	0.815574	…	0
…	…	…	…	…	0
Review N	-0.77712	-0.75144	-0.96702	-0.86927	-0.41831

Baseline Classification

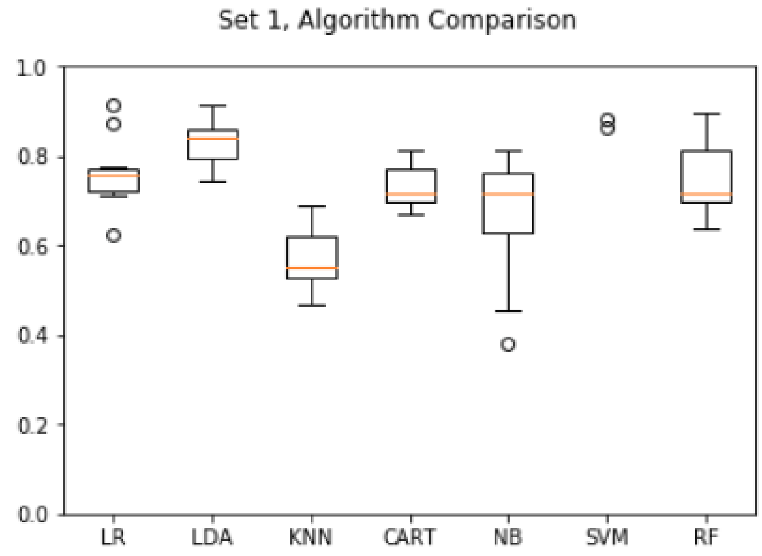
- 6 Classification Models
  - Logistic Regression
  - Linear Discriminant Analysis (LDA)
  - K Nearest Neighbor Classification
  - Decision Trees
  - Naive Bayes (NB)
  - Support Vector Classifier (SVM)
  - Random Forest (RF)

- Three sets of Features:
  - X1 = Topics Only (300 Variables)
  - X2 = Topics + Parts of Speech Metrics (300 + 4)
  - X3 = Topics + Parts of Speech Metrics + Dummy Variables (300 + 4 + 22)

Model + Feature Selection (1)

Topics Only

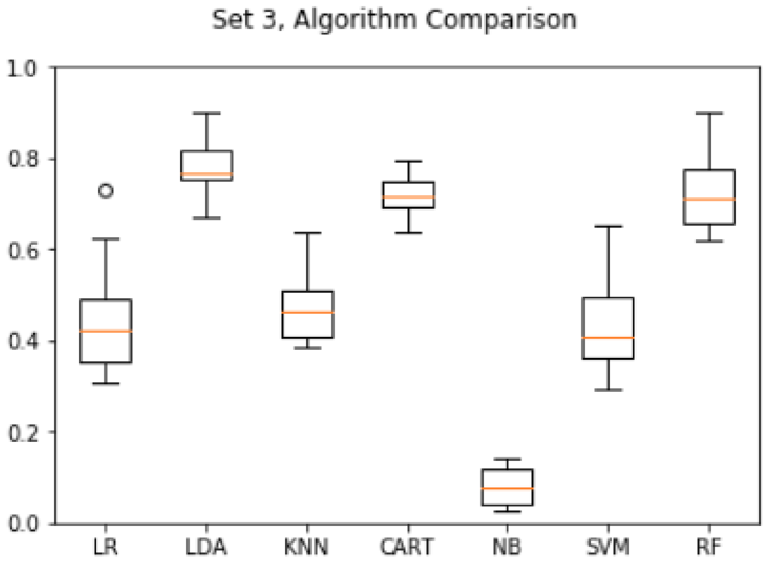
```
LR: 0.761875 (0.077533)
LDA: 0.828750 (0.048750)
KNN: 0.566250 (0.069832)
CART: 0.730000 (0.049117)
NB: 0.667500 (0.136725)
SVM: 0.174375 (0.348775)
RF: 0.748125 (0.080722)
```



LDA performs the best.

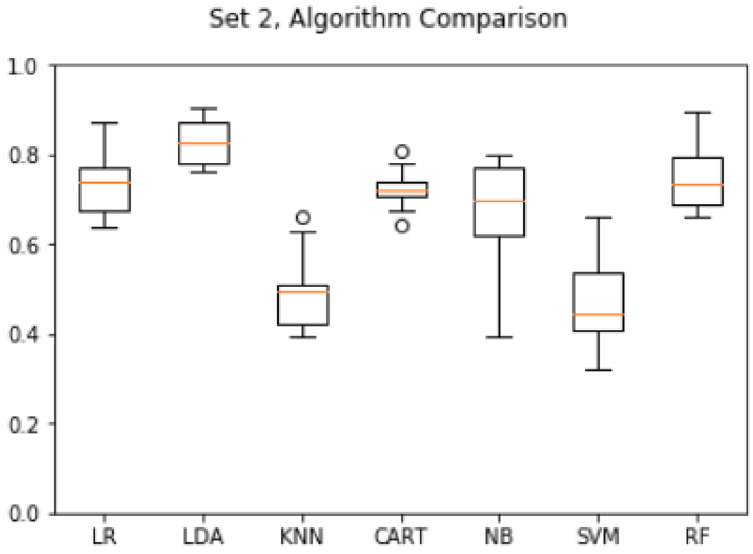
Previous + Parts of Speech Metrics

```
LR: 0.451875 (0.129845)
LDA: 0.780625 (0.062152)
KNN: 0.481875 (0.082701)
CART: 0.720000 (0.045569)
NB: 0.080625 (0.042246)
SVM: 0.436875 (0.111329)
RF: 0.728125 (0.090063)
```



Previous + Dummy Variables

```
LR: 0.736250 (0.072769)
LDA: 0.830000 (0.050683)
KNN: 0.496875 (0.086388)
CART: 0.724375 (0.044586)
NB: 0.659375 (0.135734)
SVM: 0.468125 (0.107654)
RF: 0.750625 (0.076416)
```



Going beyond baseline

Random Forest classifier because…

- Many dimensions to data
- Many parameters to be tuned
- Bonus:
    - Does well with correlated variables
    - Easily get variable importance

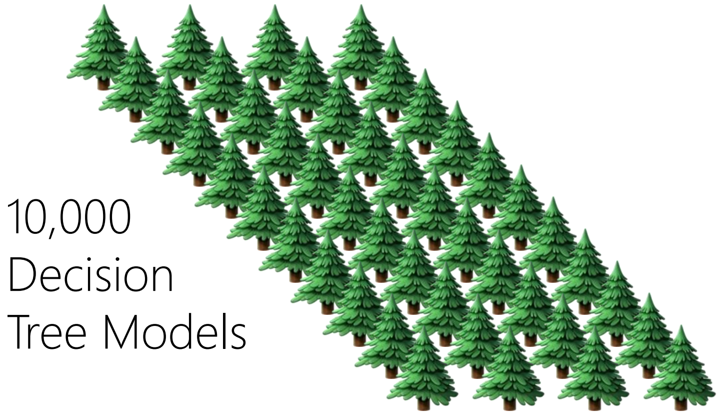
“Forest” of models

 **=**

100’s of Decision Tree models at random subsets of data

Results (1)

- Best Accuracy: 84.6%
  - 1.6 % improvement compared to LDA
  - + 20% improvement compared to untuned Random Forest model
- Best Model:
  - Max Tree Depth = 50
  - Minimum Samples per Node = 50
  - Gini Impurity
  - Of 322 variables, randomly select 18 per tree
  - 1000 Decision Trees
  - 10 folds cross validation



Results (2)

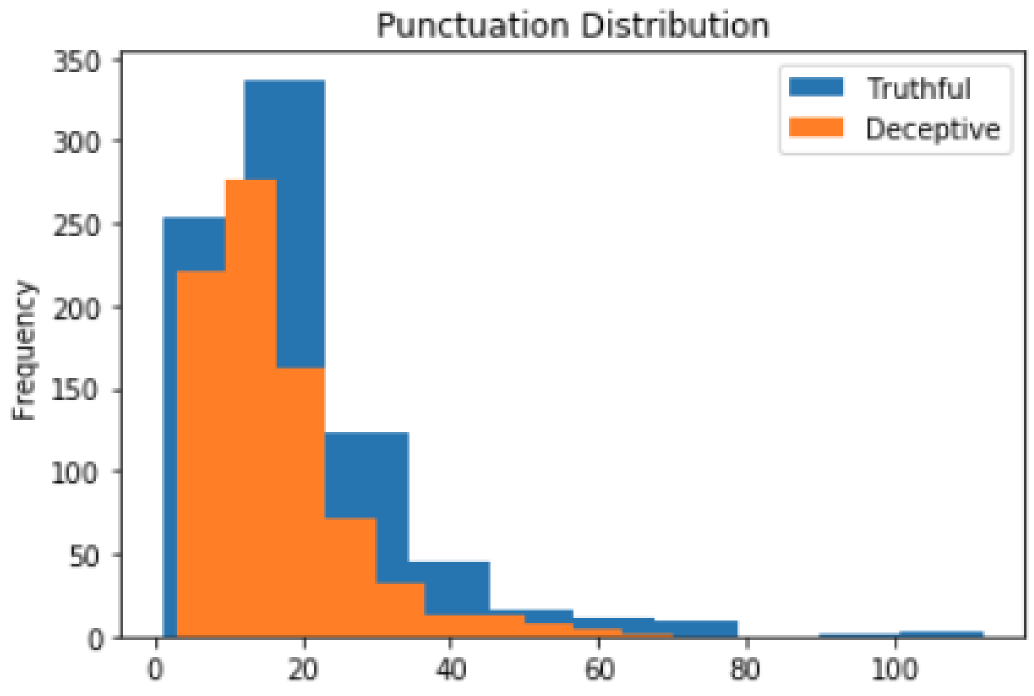
- Mean Decrease in Impurity (MDI) tells us how important a variable is
- Topic Vectors and 3 are most important
- The use of punctuation is important

	Importance	Std
2	0.266521	0.281358
6	0.051124	0.081435
3	0.047117	0.076333
4	0.035424	0.060714
12	0.024632	0.047719
15	0.024012	0.047223
punct_ct	0.019903	0.043599

	Word	Weight	topic
0	'	0.220317	2
1	great	-0.174643	2
2	comfortable	-0.129004	2
3	beautiful	-0.116429	2
4	enjoyed	-0.113951	2
5	definitely	-0.112457	2
6	When	0.110262	2
7	wonderful	-0.109365	2
8	called	0.104809	2
9	helpful	-0.104119	2
0	)	0.231953	3
1	(	0.222104	3
2	...	0.209163	3
3	$	0.204930	3
4	Hotel	-0.145028	3
5	:	0.138617	3
6	-	0.128757	3
7	Michigan	0.122856	3
8	'	0.114440	3
9	husband	-0.102649	3

Results (3)

Descriptively, there is a noticeable difference in punctuation count between truthful and deceptive reviews.



Future steps

- More data
- Vary transformation steps
- Screen predictors
- Text Processing
  - Misspellings
  - Emojis
  - N-gram
- More metadata
  - Time of post
  - User account data